A Robust Scene Descriptor based on Largest Singular Values for Cortex-Like Mechanisms

Ali Mahdavi Hormat[1], Mohammad Bagher Menhaj[2] and Vahid Rostami[1]

[1]Department of Electrical, Computer and Biomedical Engineering, Qazvin Branch
Qazvin Islamic Azad University (QIAU), Qazvin , Iran
{ali.mahdavi.hormat,vh_rostami}@qiau.ac.ir
[2]Department of Electrical Engineering Amirkabir University of Technology (AUT)
Tehran, Iran
menhaj@aut.ac.ir

Abstract— Analysis and recognition of images observed in different situations are the most important tasks of the visual cortex. Although many studies have been done in the field of computer vision and neuroscience, the underlying processes in visual cortex is not completely understood. An inspired model of the visual cortex that has lately gained attention for object recognition is HMAX, which describes a feed-forward hierarchical structure. This model shows a degree of scale and translation invariance. Other capabilities of the visual cortex are not that much sensitive against rotation and noise in color space. In this paper, we introduce a novel method to increase the degree of robustness against noise and rotation. We describe a hierarchical system that closely follows the organization of visual cortex and builds an increasingly complex and invariant feature representation with R, G, B and gray channels as inputs. Similar to the recently published methods, the phase of learning is done only in the S2 layer of HMAX structure. While in the proposed model instead of using directly the distance between the patches and the C1 units, the distance between largest singular values of the patches and the C1 units are used. These values behave indeed insensitive significantly with respect to rotation. Finally, we used the COREL datasets for experiments and showed that the proposed model have better performance than the previous HMAX models in complex visual scenes.

Keywords— *Robust Object Recognition; Visual Cortex; HMAX Model; Hierarchical System; Singular Values*

I. INTRODUCTION

The visual cortex is part of the brain for processing visual information. The discovery and analysis of cortical visual areas, major accomplishments of visual neuroscience, help us better understand how visual cortex is a critical question in neuroscience. Because humans and primates outperform the best machine vision systems with respect to almost any measure, building a system that emulates object recognition in cortex has always been an attractive but elusive and difficult goal [1]. In most cases, use of the visual neuroscience in the computer vision has been limited to early vision stages such as Stereo algorithms [2] , derivative of Gaussian as a filter and recently Gabor filters [3]. While there are some other approaches in the computer vision fully inspired and challenged by the human vision, the very first stages of processing in cells have not yet been passed [4-8].

Object recognition as one of the most advanced tasks of the visual cortex is very important for animals as well as for higher primates [9]. Although object recognition has received a great deal of attention within the field of computer vision, the underlying computational processes in the visual cortex are not completely understood and there is still a high degree of computational complexity when simulated in a computer. In short, object recognition is thought to be done by the ventral visual path from primary visual cortex, over extrastriate visual areas, to inferotemporal cortex, and to prefrontal cortex, PFC, which plays an important role in linking perception to memory [9]. This ventral visual path has a hierarchical architecture reflecting sensitivity to an increasing complexity of the preferred stimuli from simple cells in primary visual cortex to complex cells extrastriate visual areas [9, 10].

A neuroscientifically inspired model, which reflects the current understanding of the ventral visual path as a feed-forward hierarchical structure, recently gained attention in object recognition tasks is the HMAX model introduced by Reisenhuber and Poggio [11]; this model focuses more on designing simple operations inspired by the visual cortex and less on learning. Extending the work by Hubel and Weisel [12], the HMAX model, in general, possesses desirable properties observed in visual systems, and shows a significant degree of translation and scale invariance [10]. Serre et al. in [1] extended the original HMAX model to add multi-scale representations as well as more complex visual. It should be noted that the conditions under which this model functionally performs are synthetic and far from natural; however, recently the HMAX has been analyzed in natural scenes [13]. Although they are not strictly invariant to rotation [1], in this paper, we modify this model for natural scenes analysis in the rotating and noising cases by applying largest singular values and color spaces. To approve the capability of the proposed method, the COREL dataset consisting of natural images for are used in the paper.

The rest of the paper is organized as follows. Section **II** presents the structure and function of the HMAX model, Section **III** introduces the proposed model by using the RGB and gray level color spaces and largest singular values in the learning phase. Section IV presents the simulation results from the experiment carried out to test and compare these models under different conditions. Finally, section V concludes the paper.

II. HMAX's Structure and function

The algorithm to implement this model performed for number of scale bands that each scale band determines the size of the filters employed and the number of units pooled (size of a local area for MAX operations into one scale band). This will be explained below.

A. Standard Model

According to Fig.1 in its simplest form, the model consists of four layers of computational units, where the simple S units alternate with complex C units. The S units increase selectivity by tuning function [11, 12]. The C units pool their inputs through a maximum (MAX) operation to increase invariance in the spatial and scale. The model consists of several properties of cells along the ventral stream of visual cortex [14]. For instance, operations at the complex cells are similar to some behaviors of complex cells in V1 [15] and V4 [16].
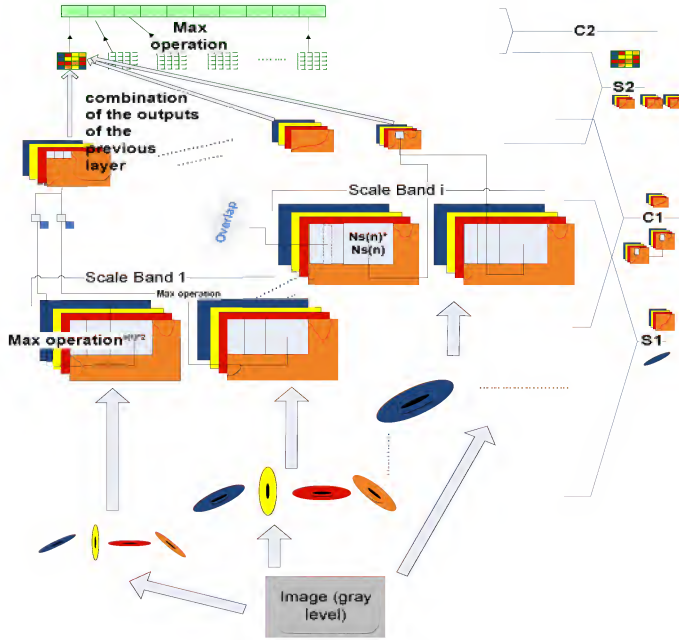

Fig.1. Illustrate the standard HMAX model structure. S1 layers consists of simple Gabor filters at several scales and orientation which indicated by ellipses with various of size and color (each color to mean one orientation). C1 pools the filter output spatially (gray squares with a size of Ns) and across adjacent scales into scale band. S2 tuned to conjunctions of the orientations. C2 provides further spatial and scale invariance. The C2 outputs are directly fed to a classifier .

B. Modified HMAX model

This model is based on occurs learning in the all stages in the visual cortex. In the S2 layer is used of the learning to obtain a good performance in the robust object recognition [1] or combined with other features for scene classification [17]. Each S2 unit response depends in a Gaussian-like way on the Euclidean distance between a new input and a stored prototype (see Fig.2). That is, for an image patch X from the previous C1 layer at a particular scale, the response r of the corresponding S2 unit is given by

$$r = \exp(-\beta \|X - P_i\|^2) \qquad (1)$$

Where β defines the sharpness of the tuning and P_i is one of the centers (look like to the RBF units), extracted randomly.

III. The proposed model

The core of the proposed model focuses on increasing process tolerance against rotation and noise during the phase maximizing of color spaces and intensities. In this model, Starting point for the proposed model is decomposes an input image to four channels (R, G , B and intensity) and do Max operation over them. So, in the S1 layer, we apply the Gabor filters (by using (2) and (3)), adjust the filter parameters, i.e. orientation θ, effective width σ, wavelength λ. S fully discussed in [18],we then apply some of those filters given in Table.I to stimuli commonly used ones to probe V1 neurons and to remove filters that are incompatible. We use also these filters to form a pyramid of scales in the o- orientations. These filters grouped in specified scale bands.

$$G(x, y) = \exp(-\frac{(x_0^2 + \gamma^2 y_0^2)}{2\sigma^2}) \times \cos(\frac{2\pi}{\lambda} x_0) \quad s.t. \quad (2)$$

$$\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix}^T \times \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \qquad (3)$$
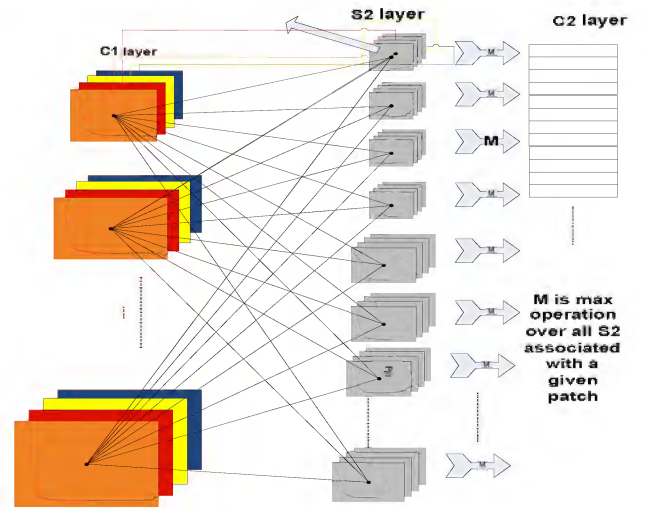

Fig.2. The N patches in the S2 layer with C1 format. Each orientation in the patch is matched to the corresponding orientation in C1. The results are one image per C1 band and patch. The C2 values are computed by taking a MAX over all S2 associated with a given patch. Thus, the C2 response has length N.

The C1 units pooled using a maximum operation of a local area ($Ns \times Ns$) of the S1 units from the same orientation and scale band. This pooling increases the tolerance against the shift. That is the response r of a complex unit (by using (4)) corresponding to the response of the strongest of its cells ($x_1, x_2, ..., x_m$) from the previous S1 layer.

$$r = \max_{j=1,...,m} x_j \qquad (4)$$

For instance, we have several maps in a scale band, filters with various sizes. The maps have the same dimensionality. But they are the outputs of different filters. For pooling, one measurement is a maximum operation over the same spatial neighborhood of the filter's output.

It is also proposed that for each S2 units, the largest singular values of both patches and C1 units are computed. The prototypes P or patches with size of $n \times n$ and o-orientations are randomly extracted from the C1 layers of training images and then stored. There are two reasons for using largest singular values as illustrated by the following two examples:

Example 1: Tolerance against rotations

We assume that the patch is $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$, therefore

$P_1 = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$ and $A_1 = P \times P^T$ or A_1 is equal to $\begin{bmatrix} 14 & 32 & 50 \\ 32 & 77 & 122 \\ 50 & 122 & 194 \end{bmatrix}$

Now if we rotate P_1 to 90°, P_2 is

$P_2 = \begin{bmatrix} 3 & 6 & 9 \\ 2 & 5 & 8 \\ 1 & 4 & 7 \end{bmatrix}$, A_2 is equal to $\begin{bmatrix} 126 & 108 & 90 \\ 108 & 93 & 78 \\ 90 & 78 & 66 \end{bmatrix}$ and then for

the largest eigenvalue of A_1 and A_2 we have:

$$\begin{bmatrix} 14 & 32 & 50 \\ 32 & 77 & 122 \\ 50 & 122 & 194 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \longrightarrow \begin{cases} (14-\lambda)x_1 + 32x_2 + 50x_3 = 0 \\ 32x_1 + (77-\lambda)x_2 + 122x_3 = 0 \\ 50x_1 + 122x_2 + (194-\lambda)x_3 = 0 \end{cases} \qquad (5)$$

$$\begin{vmatrix} 14-\lambda & 32 & 50 \\ 32 & 77-\lambda & 122 \\ 50 & 122 & 194-\lambda \end{vmatrix} = 0 \longrightarrow \lambda = \lambda_i, i = 1,2,3,... \quad, \lambda_1 > \lambda_2 > ... \qquad (6)$$

$\lambda_{A1} = \lambda_1 = 283.8586$

$$\begin{bmatrix} 126 & 108 & 90 \\ 108 & 93 & 78 \\ 90 & 78 & 66 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \longrightarrow \begin{cases} (126-\lambda)x_1 + 108x_2 + 90x_3 = 0 \\ 108x_1 + (93-\lambda)x_2 + 78x_3 = 0 \\ 90x_1 + 78x_2 + (66-\lambda)x_3 = 0 \end{cases} \qquad (7)$$

$$\begin{vmatrix} 126-\lambda & 108 & 90 \\ 108 & 93-\lambda & 78 \\ 90 & 78 & 66-\lambda \end{vmatrix} = 0 \longrightarrow \lambda = \lambda_i, i = 1,2,3,... \quad, \lambda_1 > \lambda_2 > ... \qquad (8)$$

$\lambda_{A2} = \lambda_1 = 283.8586$

$\lambda_{A1} = \lambda_{A2}$, so that root square of them as singular values is $\delta_{p_1} = \delta_{p_2} = 16.841$

Example 2: No use of the patch matrices

$P_1 = \begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix}$ and $P_2 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

$mean_{P_1} = \dfrac{0+0+0+4}{4}$, $mean_{P_2} = \dfrac{1+1+1+1}{4}$ that $mean_{P_1} = mean_{P_2}$ but $P_1 \neq P_2$

$\delta_{P_1} = 4, \delta_{P_2} = 2$, so that $\delta_{P_1} \neq \delta_{P_2}$ (δ is the singular value).

For each C1 image (output of a band), compute:

$$Y_{\theta_i} = \exp(-\xi \left\| \delta(x_{\theta_i}) - \delta(pj_{\theta_i}) \right\|^2) \qquad (9)$$

Which P_i is patch extracted randomly. Then in the C2 layer, for each unit we compute:

$$F_j = \sum_i^o Y_{\theta_i} \qquad (10)$$

The size of F_j is equal to the number of patches. So, we have a complex feature vector or F that its tolerance against rotate, noise, size and shift.

IV. OVERALL ACCURACY RESULTS FROM ROTATING AND NOISE CASES

We considered four orientations 0°, 45°, 90°, 135° and arranged the S_1 filters to form a pyramid of scales. According to Table.I we used the 4 scale bands which each consists two adjacent filter sizes (there are 4 scale band for a total of 8 filter sizes). In addition, Table.I for each of the scale bands determined the size of S_1 neighborhood or $Ns \times Ns$. For all the Gabor filters used $\gamma = 0.3$, other parameters these filters defined in Table.I, for each the size of them. We considered 20 patches with sizes of 10 which these randomly extracted from the training images.

We extracted 50 complex feature vector of positive samples and 50 complex feature vector of negative samples as train set, which these selected randomly. The SVM trained over this set. For evaluate the proposed method in noisy cases, the images relate to this set changed by pepper and salt noise with 0.1 value. Then, test set consists of the complex feature vectors obtained from the noising images. According to Table.II, for five iterations (in each iteration again the training images are selected) the test sets are given to the trained SVMs and the performance based on average mean square errors (MSE) is measured.

Again similar to the noise cases, we extracted 50 complex feature vectors of positive samples and 50 complex feature vector of negative samples as train set selected randomly. The SVM trained over this set. To evaluate the proposed method in rotating cases, the images relate to this set changed by rotation 1 to 180, randomly. Then, the test sets given to the trained SVMs and measurement performance based on average mean square errors (Table.III).

TABLE. I. Parameters used in our implementation for the S1 and C1 layers.

	For C1 layer		For S1 layer		
Scale band	Ns	Overlap	Filter size	σ	λ
1	4	2	7	2.8	3.5
			9	3.6	4.6
2	6	3	11	4.5	5.6
			15	5.4	6.8
3	8	4	17	6.3	7.9
			19	7.3	9.1
4	12	6	21	8.2	10.3
			23	9.2	11.5

TABLE. II. Comparison of the our proposed model vs. other HMAX models by average mean square errors (MSEs) over 5 rounds of Test at noise cases on the COREL date set.

Image category	The our proposed model	The modified HMAX [1]	The standard HMAX [11]
Africa people	76%	40%	52%
Beaches	78%	59%	45%
Buildings	82%	65%	58%
Buses	82%	65%	46%
Dinosaurs	78%	79%	43%
Elephants	77%	62%	49%
Flowers	69%	50%	53%
Horses	70%	66%	36%
Mountains	82%	60%	51%
Foods	73%	49%	54%
Average	76%	59%	48%

Because the previous models used the MAX operation for choosing components values of the feature vector (F), there is a high probability that they choose the noisy values. This means that one noisy pixel in the local area for pooling has the maximum value, this value may be transmitted by MAX operation from the S1 layer to C2 layer and in the last, it is selected as a component of F. However, the proposed model in the last layer uses the relationship between all the values (singular value). The low percentage of the Standard HMAX model [11] indicates disability of the this model for analysis and description of complex visual scenes.

TABLE. III. Comparison of the our proposed model vs. other HMAX models by Average mean square errors (MSEs) over 5 rounds of Test at rotating cases on the COREL date set.

Image category	The our proposed model	The modified HMAX [1]	The standard HMAX [11]
Africa people	75%	70%	61%
Beaches	78%	57%	53%
Buildings	75%	72%	73%
Buses	91%	62%	55%
Dinosaurs	80%	52%	63%
Elephants	69%	47%	54%
Flowers	70%	53%	60%
Horses	73%	61%	64%
Mountains	71%	58%	62%
Foods	74%	73%	71%
Average	75%	60%	61%

It can be seen from Table.II and Table.III, among all images in the 10 categories, our model achieves the best overall performance in describing Images. Similar to the visual system of the brain, this model inspired of the visual cortex has capability tolerance against noise and rotation. With attention to using the color in visual cortex into the V1 area, we used the color channels. Therefore, the obtained features have some advantages besides the aforementioned weaknesses. A weakness can be used from the small number of patches to reduce time complexity and the image information lost. To better illustrate the performance of our method in recognizing the images corrupted by noise and rotation, we used the images of Africa and beaches categories and plotted the ROC curves of two different models in the 5 iterations for the noisy cases (see Fig.4) and rotating cases (see Fig.5), respectively.
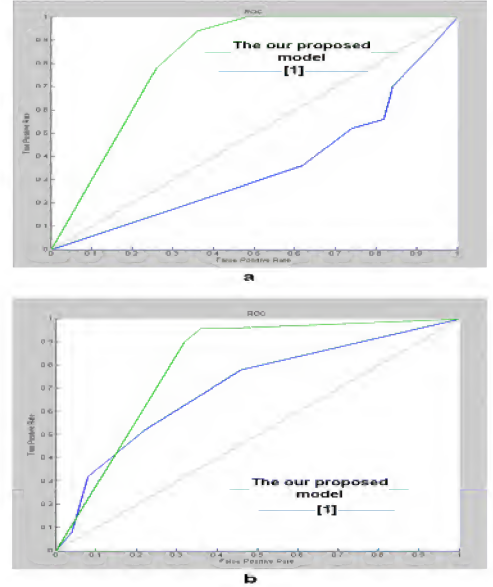

Fig.4. Comparison of the ROC curves of two different models in the noising cases: (a) Africa (b) Beaches categories in 5 rounds
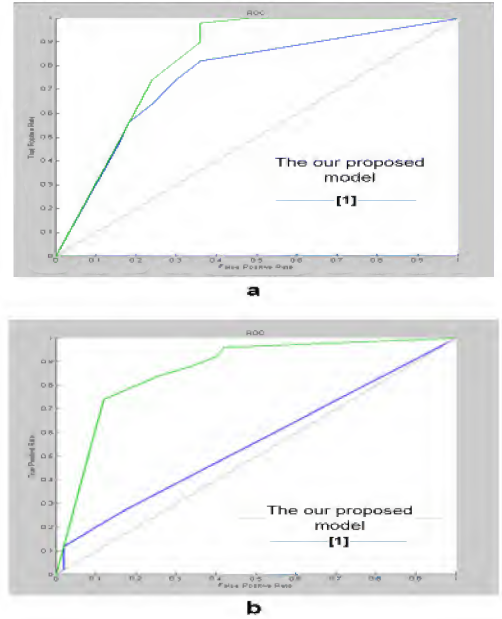

Fig.5. Comparison of the ROC curves of two different models in the rotating cases: (a) Africa (b) Beaches categories in 5 rounds

A. Speed

The time complexity of the singular value decomposition (SVD) is $O(m^2n+n^2)+k$ which m and n are the dimensions of the patch (k is computed as the SVD of S2's output) . We applied the square patches, the time complexity is $O(n^3+n^2)$. On the other hand, the time complexity for Euclidean distance discussed in [1] was used with a complexity of $O(k_jn_j^2)$ in which k is the magnitude S2's output to patch of size 4 patches, j=1, 2, 3, 4.

If we assume k and k_j equal to 1, for this model n=5 and number of the patches becomes 20, q=20 , then

$T_1 = q(\,n^3+n^2\,)=20(5^3+5^2)=20\times150=3000$

And for the modified HMAX [1], use $n_1=4$, $n_2=8$, $n_3=12$, $n_4=16$ and numbers of the patches $q=250$ and obtain

$T_2=q(\,n_1^2+n_2^2+n_3^2+n_4^2\,)=250(4^2+8^2+12^2+16^2)=250(480)=120000$

Therefore we have $T_1 \ll T_2$.

Table.IV lists only in average the time of creating feature vector (in Matlab 7 on a Pentium(R) Dual Core T4200+PC running the windows XP operating system with 2G memory) for one train sample. In this case (in tolerance against rotation and noise), we have seen the proposed model possesses a higher speed than that of the modified HMAX model [1].

For the proposed model and other models the training time is mainly spent in two aspects: (1) to generate a collection of feature vector to construct a training set, (2) to do SVM training over feature vectors of images. Because, the dimension of the previous model is high, it incur significantly higher computational cost than that of the proposed model as shown in Table.V.

CONCLUSIONS AND FUTURE WORK

In this paper, we presented a new model with a framework similar to the previous models that inspired by visual cortex. The proposed model, in which the capabilities of visual cortex for scene recognition for noisy and complex environment are preserved, employed color channels and largest singular values of patches in the HMAX structure. The proposed model first computes a set of rotating, noise and shift invariant, scale- free features from a training set of images. Then, a standard discriminative classifier on the vector of features obtained from the input images is applied. Our approach exhibited higher performance on the COREL images set in compared to other models. A weakness of the proposed model is its high time complexity for computing singular values; to remedy this, and we had to use a smaller number of patches. This will also be pursued in our future research.

TABLE. IV. Comparison time of the our proposed model vs. other model for creating one feature vector (in seconds)

The our proposed model	The modified HMAX model [1]
1.29172316 seconds	9.89557172 seconds

TABLE. V. Comparison time of the our proposed model vs. other model for creating training set and SVM training (in seconds)

Steps	The our proposed model	The modified HMAX model [1]
To generate training set	64.586158 seconds	494.778586 seconds
For SVM training	0.03346 seconds	0.03590 seconds

REFERENCES

[1] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber and T. Poggio, "Robust object recognition with cortex-like mechanisms," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 29, pp. 411-426, 2007.

[2] W. E. L. Grimson, "A computer implementation of a theory of human stereo vision," Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, pp. 217-253, 1981.

[3] J. P. Jones and L. A. Palmer, "An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex," Journal of Neurophysiology, vol. 58, pp. 1233-1258, 1987.

[4] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," Biological cybernetics, vol. 36, pp. 193-202, 1980.

[5] B. W. Mel, "SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition," Neural computation, vol. 9, pp. 777-804, 1997.

[6] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, pp. 2278-2324, 1998.

[7] H. Wersing and E. Körner, "Learning optimized features for hierarchical models of invariant object recognition," Neural computation, vol. 15, pp. 1559-1588, 2003.

[8] B. Yang, L. Zhou and Z. Deng, "C-HMAX: Artificial cognitive model inspired by the color vision mechanism of the human brain," Tsinghua Science and Technology, vol. 18, pp. 51-56, 2013.

[9] M. Riesenhuber and T. Poggio, "How visual cortex recognizes objects: The tale of the standard model," in: The Visual Neurosciences, MIT Press, 2003.

[10] N. Logothetis, J. Pauls, H. Bulthoff and T. Poggio, "View-dependent object recognition by monkeys," Current biology, vol. 4, pp. 401-414, 1994.

[11] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," Nature neuroscience, vol. 2, pp. 1019-1025, 1999.

[12] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," The Journal of physiology, vol. 160, p. 106, 1962.

[13] D. Walther, L. Itti, M. Riesenhuber, T. Poggio and C. Koch, "Attentional selection for object recognition—a gentle way," in Biologically Motivated Computer Vision, 2002, pp. 472-479.

[14] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman and T. Poggio, " A Theory of Object Recognition: Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex," AI Memo 2005-036/CBCL Memo 259, Massachusetts Inst. of Technology, Cambridge2005.

[15] I. Lampl, D. Ferster, T. Poggio and M. Riesenhuber, "Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex," Journal of Neurophysiology, vol. 92, pp. 2704-2713, 2004.

[16] T. J. Gawne and J. M. Martin, "Responses of primate visual cortical neurons to stimuli presented by flash, saccade, blink, and external darkening," Journal of Neurophysiology, vol. 88, pp. 2178-2186, 2002.

[17] A. Borji and L. Itti, "Scene classification with a sparse set of salient regions," in Robotics and Automation (ICRA), 2011 IEEE International Conference on, 2011, pp. 1902-1908.

[18] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat," Journal of Neurophysiology, 1965.